图书情报工作
第62卷 第14期 2018年7月

基于主题图的数字档案标注系统资源聚合研究*

张云中 冯双双
上海大学图书情报档案系 上海 200444

摘要：[**目的/意义**]针对社会化标注系统用于数字档案资源组织后带来的资源检索与导航问题，提出依托主题图的数字档案资源聚合模型，以期提高数字档案资源检索效率并建立有序的可视化导航。[**方法/过程**]在剖析利用主题图实现社会化标注系统资源聚合相关研究现状的基础上，构建数字档案领域基于主题图的资源聚合模型，给出利用社会网络分析和形式概念分析析取数字档案资源主题图的主题类型、关联关系及资源指引三要素的体系化解决方案，从而实现数字档案标注系统资源聚合。[**结果/结论**]以NARA数字档案标注系统中的"Women at War"话题为例，利用所提方法并结合Ontopia工具实现目标话题的数字档案资源聚合，有效提高系统中数字档案资源的检索效率和导航效果。

关键词：社会化标注系统 主题图 资源聚合 数字档案
分类号：G254.11
DOI：10.13266/j.issn.0252-3116.2018.14.014

利用社会化标注系统组织数字档案资源是近年来档案实践领域兴起的资源组织新方案，美国NARA（National Archives and Records Administration）国家档案与文件署的公民档案标注系统即是此类实践活动中的最佳示范，受到了档案领域学者及档案爱好者的广泛青睐。采用社会化标注的方式组织数字档案资源优势明显，诸如集体智慧、更新及时、自由灵活、用户体验感强等，但标签语义规范性差、标签结构扁平化等固有缺陷导致的检索与导航问题也接踵而至，借助其他知识组织方法对标签语义进行优化，实现基于语义的数字档案资源聚合成为学者们破解该难题的共识[1]。如是，主题图即是该难题的一解。主题图的核心三要素是主题（topic）、关联（association）和资源出处（occurrence），通过精准描述主题及主题之间、主题与资源之间的形式化语义关系，可形成直观的可视化导航图，其规范化、形式化、准确性、可视化等优点与标签形成了鲜明的互补特色，可以推测两者的结合是解决数字档案资源社会化标注系统缺陷行之有效的方案，而实质上已有诸多学者照此思路展开了一些有特色的研究[2-4]，尝试从不同角度建立标签与主题图之间的映射关系。本研究旨在参考国内外既有研究的基础上，重构社会化标注系统三元组向主题图三要素之间的映射方案，尝试采用社会网络分析、形式概念分析等量化工具，使得建立映射的过程更为科学、严谨且尽量弱化主观性，以期建立的基于主题图的数字档案资源聚合结果能描述更精准的语义和展示更精确的导航。

1 文献述评

国内外少有利用主题图解决数字档案标注系统资源聚合问题的研究，但在数字图书馆资源聚合、学术博客资源聚合等相似问题上却形成了一批可以借鉴的研究成果，这些文献关注的核心问题有如下两点：①主题图能否拨开标签云的天空？这是一个源自TMRA2007会议的一个形象比喻[5]，其本质是探讨主题图和社会化标注系统结合的可行性。该类研究多从社会认知、技术实现等角度探讨二者的结合问题，例如D. Hendel[6]从社会和认知角度对主题图在社会网站中的应用进行考察，肯定了主题图与标签结合的可行性；陈婷[2]则从知识组织、语义关联和技术互补等角度肯定

* 本文系国家哲学社会科学基金项目"基于形式概念分析的社会化标注系统语义发现与语义映射研究"（项目编号：16CTQ023）研究成果之一。
作者简介：张云中（ORCID：0000-0002-7323-2561），副教授，博士，硕士生导师，E-mail：zhang-yun-zhong@126.com；冯双双（ORCID：0000-0002-6397-8714），硕士研究生。
收稿日期：2018-01-11 修回日期：2018-03-16 本文起止页码：116-124 本文责任编辑：王传清

了标签与主题图结合的可行性。国内外学者均普遍认可采用二者结合的方式优化数字资源组织。②主题图如何拨开标签云的天空以实现资源聚合？关于如何利用标签主题图结合实现社会化标注系统资源组织及聚合，国内外学者采用的方法及应用的领域呈现多样化：K. Fujimura[7]在博客导航系统中采用数据挖掘技术，用主题图对大规模的标签云进行整理和序化以揭示标签关系；熊回香[8]和邓敏[9]在标签分类的基础上抽取主题类型并主观赋予主题关联以实现豆瓣电影标注系统中的主题图构建；夏立新等[10]在知识专家学术社区构建领域介绍了 Fuzzy 标注系统中利用主题图实现标签互联的方案；项兴彬[5]采用与已有文献[8]和[9]类同的方式对工程建设中的标签资源进行了主题类型、关联、资源指引定义，建立起新的标签主题图的资源组织模型。

综上，既有研究提供了非常有价值的求解框架，给出了利用主题图实现各个领域信息资源聚合的一般性解决方案。但既有研究也仍然有尚未解决好的问题，主要包括：①主题类型的遴选多采用标签分类基础上参考既有分类标准自定义主题类型，由此方式产生的主题类型语义粒度粗放；②主题关联关系的确定多依赖主观，缺少客观的分析过程及参照标准；③资源指引往往被忽视，资源的聚合过程未被凸显；④主题类型和主题的定义侧重于对信息资源外部特征的描述，忽视了对资源内容特征的揭示。上述 4 个问题正是本研究拟解决的关键所在。

2　基于主题图的社会化标注系统资源聚合的模型构建

基于主题图的社会化标注系统资源聚合的本质，是用主题图形态重新组织原来以标签形态展示的社会化标注系统资源，因此问题的关键可抽象为建立社会化标注系统{标签集，资源集，标签－资源关系集}集合向主题图{主题，关系，资源指引}之间的映射。国内外同类研究建立此映射的一般思路是将标签分类进而映射为主题类型及主题，主观性分类产生的标签间关系映射为主题关系，资源的 URI 标识映射为资源指引。本研究在文献述评中也提及了目前这种主流映射方式的局限，为了弥补上述局限，本研究拟定了新的映射方案：①采用先聚类再分类的处理方式，以自底向上的聚类代替主观性自顶向下的分类，完成标签向主题类型及主题的映射，使得主题划分更科学，语义粒度更细致。②采用以概念关系分析的客观方式提取主题间关系，以代替人为自定义的主题关系，完成标签关系向主题关系的映射，使得类属、相关等关系的确立更客观。③给出详尽的聚合资源指引方案，完成资源集向资源指引的映射，使得资源能以聚合的形式展示和导航。

为更清晰地说明该方案思路和任务，本研究构建了基于主题图的社会化标注系统资源聚合模型，该模型主要涵盖数据处理、数据分析、结果展示 3 个模块，如图 1 所示：

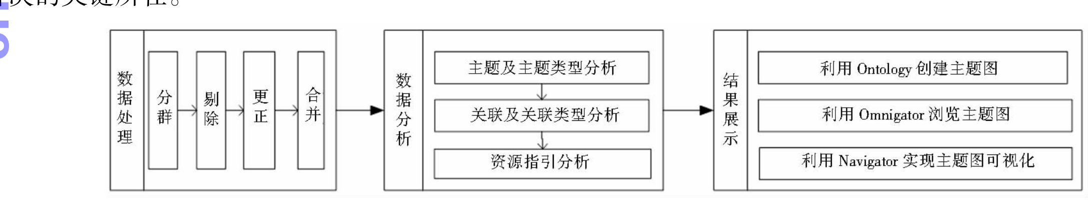


图 1　基于主题图的社会化标注系统资源聚合模型

2.1　数据处理模块

数据处理模块旨在将社会化标注系统中抽取出的{资源集，标签集，标签－资源关系集}展开预处理，为数据分析模块奠定基础，数据预处理的关键环节包括分群、剔除、更正和合并：①分群：本研究侧重从资源内容特征的角度建立主题图，因而需先将标签按照描述资源外部特征和内容特征进行区分，描述资源内容特征的标签集是本研究着重关注处理的数据对象；②剔除：将无标签描述的资源及一些无意义或无效的标签去除；③更正：错拼、错写的标签修改；④合并：英文缩写、单复数、大小写、人名地名的合并。

2.2　数据分析模块

数据分析模块旨在获取的精炼数据集基础上利用特定的分析方法展开主题及主题类型分析、关联关系分析和资源指引分析，建立{资源集，标签集，标签－资源关系集}向{主题类型集，主题关系集，资源指引集}的映射关系，从而实现基于主题图的社会化标注系统资源聚合。

2.2.1　主题及主题类型分析　主题是主题图中描述知识的基本构成单元，是对客观事物的抽象化描述。主题可以划分为群，谓之主题类型，一个主题可以归属

于一个以上主题类型。主题类型不仅可以从资源外部特征中提取，还可以从资源内容特征中抽象而出。社会化标注系统中的标签集兼顾对资源内外部特征的描述，因而，从中遴选和提取主题及主题类型是依托主题图实现社会化标注系统资源聚合的不二选择。

本研究侧重从资源内容特征的角度建立主题图，故重点以揭示资源内容特征的精炼“标签 – 资源”数据集为数据源，采用“先聚类再分类”的处理思想，通过构建高频标签共现矩阵，进而利用社会网络分析工具判定标签间语义距离之远近亲疏，据此将标签集聚类为若干标签群，借以发现主题类型，见图 2。本研究中高频标签的遴选与文献计量中高频关键词遴选方案异曲同工，在此不予赘述。另外，为保障聚合分析的正确性，本研究采用两种聚合工具——NetDraw 和 NodeXL 互为印证。综上，所遴选的高频标签可视为主题，聚合而出的标签群冠名后即为主题类型。

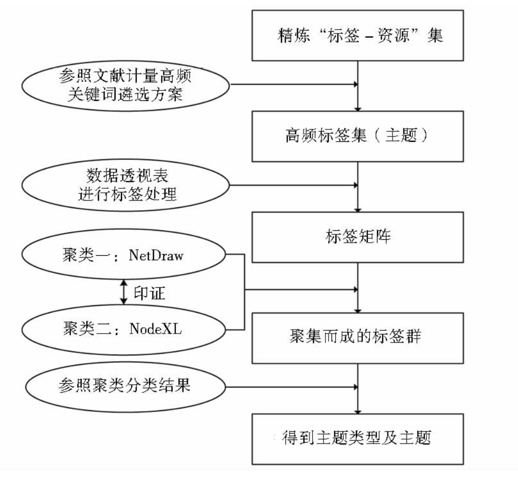


图 2　主题及主题类型发现

2.2.2　关联及关联类型的分析　关联是揭示主题之间语义关系并连接相关主题形成完整的语义网络的关键要素，其设立以参考专家经验自定义语义关联为多见，但难脱主观之嫌。为此，本研究采用应用数学中的形式概念分析方法来识别判定主题间语义关系，使得分析过程更为客观。形式概念分析理论以数学化的形式概念为基本知识单元，以形式背景描述概念内涵外延间关联，以概念格中节点的泛化和例化抽象表示概念、属性、实例间的各类关系，进而描述精准的语义关系，适合解决本研究中的主题关系求解问题。

接上步，以聚类后选定的一个主题类型（即聚类所得的某个标签群）及其所含主题（即标签群中所含标签）为数据源，将该主题类型所蕴含数据按照“标签 – 资源”的二元关系装载入形式背景，进而转换为概念格得到标签间的层级关系，其本质是利用聚类算法将具有相同主题的资源进行聚集，使得主题类型中的主题呈现出从无序到有序的结构。假定图 3 中所示的形式背景由某主题类型所含数据装载而得，标签 i 为形式概念的内涵，资源 j 为概念的外延，以“×”代表标签 – 资源的对应关系，可将其转换为图 3 所示的概念格。该概念格中，节点 1 与节点 2 为概念属分关系，可以此为依据推理主题 A 与主题 D 为属分关系；类似地，节点 2 与节点 3 的交集为节点 4，二者为相关关系，可据此主题 B 与主题 D 为相关关系。因而，以形式概念分析为工具，可从资源内容特征的角度揭示包含、属分、相关等多种关联类型，此即是关联及关联类型分析。

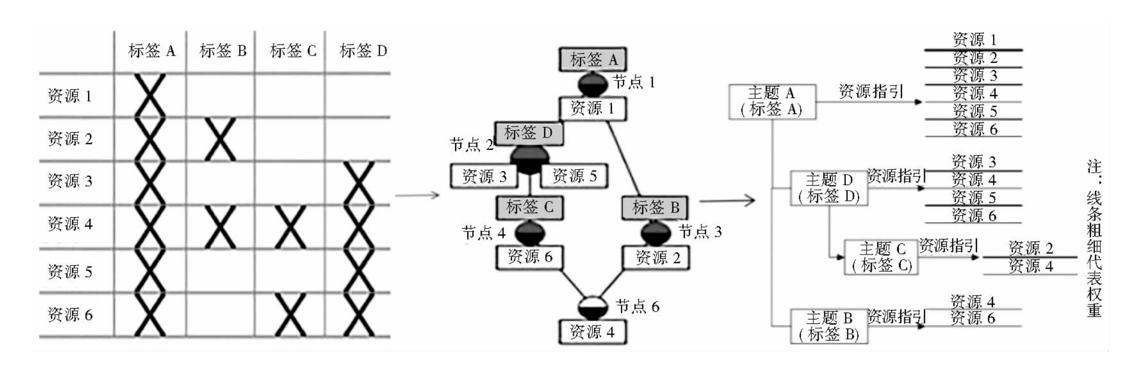


图 3　关联、关联关系分析和资源指引分析

2.2.3　资源指引分析　资源指引是指确立主题及主题关联之后，在相应的主题下链接资源实体的过程。资源实体是独立于主题图外的描述特定主题的网页、图片、数据、文本、视频等各种资源，可为社会化标注系统中的固有资源，亦可通过拓展链接社会化标注系统之外的资源，其一般采用 HTML、URI、Number、Datetime、String、Image 等资源指引类型来界定主题类型和资源实体的关系。本研究着重关注对 STS 资源的权重

指引，强调通过资源指引实现基于主题图的资源聚合与导航。以图3右半部分为例，确立主题关系后，不难发现作为外延的资源是存在层级关系的，依据形式概念分析理论，可将其解释为概念外延的逆向继承性。在资源检索时，这种逆向继承性可以用以描述所获资源的权重，供排序和优先推荐之用。例如，未建立聚合式资源指引前，检索标签D，可以检到资源3、资源4、资源5和资源6，各资源权重相同；采用聚合式资源指引后，仍可检索到上述资源集，但资源权重有别，资源3和资源5的排序和推荐应优先于由逆向继承产生的资源4和资源6。

2.3 结果展示模块

结果展示模块旨结合特定的主题图构建工具，将3类分析结果使用主题图表示工具描述和展示给用户，以最终实现基于主题图的社会化标注系统资源聚合。目前较为主流的主题图构建工具有TM4J、tinyTIM、XTM4XMLDB和Ontopia。在此模块中，本研究选用学者们使用频率相对较高的主题图工具Ontopia来"描述"3类分析结果从而构建主题图：对主题类型及其所含主题的"描述"可用Ontopia中的Topic Types模块实现；对关联及关联关系的"描述"可用Ontopia中的Association Types模块创建，可描述的关联关系涵盖包含关系、属分关系、相关关系等；对资源指引的"描述"可用Ontopia中的Occurrence Types模块创建与对应主题相关的资源属性、资源类型和资源链接；基于主题图的资源聚合与导航结果可通过Omnigator的主页面来展示，用户可直接浏览主题、关联关系、资源指引及其指引所给出的链接，并通过点击链接，到达相应的信息资源，从而将内部的主题、关联等和信息资源联系起来。基于主题图的资源聚合与导航结果亦可通过Navigator来实现主题图可视化，将主题与主题之间的关系形成一个用以表达语义的网状结构。通过对主题所表示出的关联进行追踪查询，可以了解更多相关资源，提高检索系统的查全率。

3 例证研究：基于主题图的NARA数字档案资源聚合方案

3.1 数据获取与清洗

本研究主要以NARA数字档案馆中Citizen Archivist Dashboard板块的tagging missions英文标签资源作为数据源，用八爪鱼采集器抓取其中一个tagging mission"Women at War"下用户对其381件档案标注的标签，截至2017年9月26日共计1 836个，本研究将获取的标签导入Excel表格中使用筛选、替换、查错、排序等功能进行分群、剔除、更正、合并等人工清洗操作，清洗规则见表1，得到最终的档案记录数是248条，标签数是1 695个。

表1 数据清洗规则示例

清洗顺序及依据	示例	操作
1 无标签的资源	Nationalarchives identifier＝44266358	删除44 266 358这条档案
2 管理员标签无实质含义标签	amam-ts1	删除仅有amam-ts1标签的档案
3 错拼	Wolrd War II	修正为World War II
4 缩写、单复数、大小写	Women's Army Corps＝WAC Women＝woman	合并为Women's Army Corps 合并为women
5 人名、地名合并整理	Women Marines、Marines	合并为Women Marines
最终结果记录数、标签数	记录数:248　　标签数:1 695	

标签清洗整理后，借鉴文献计量学中高频关键词选取的思路提取出高频标签见表2。词频筛选规则为：先取词频2以上的标签共计92个，词频中位数为4，然后将词频4及以上的标签作为高频标签。

根据表3给出的高频标签，可得到标签53个，总标签词频数为702，然后使用excel里的数据透视表，得出53＊53的共现矩阵（限于篇幅，只给出部分，见表3）。

标签共现矩阵中，每个数字对应的是其行标签与列标签的共现次数，数字大小代表两个标签的关联关系的强弱。共现标签间的关联关系也间接体现了被其标注的档案资源的关联关系，通过对标签及标签间关系的分析，可实现基于主题图的NARA数字档案资源聚合。

3.2 数据分析

3.2.1 主题及主题类型分析　本步骤旨在通过聚类分析判定标签关系的强弱进而发现主题类型及其所包含主题，为确保聚类结果的精准性，本研究采用NetDraw和NodeXL两种聚类工具分别聚类、相互印证。

（1）将前文所得53＊53的共现矩阵导入NetDraw中，通过"分析（analysis）"菜单中的"中心性测量（centrality measures）"功能，使用"Degree（描述特定节点到其他节点的直接联结数目）"作为测量要素，对所选高频标签在网络中的中心地位及标签间的语义亲疏展开

表 2 高频标签筛选

序号	标签	词频	序号	标签	词频
1	women	110	28	World War II Posters	7
2	World War II	80	29	food	6
3	World War I	67	30	United States Navy	6
4	women war workers	45	31	Vassar College	6
5	posters	25	32	women in war	6
6	nurses	20	33	Women's Bureau	6
7	united states army	17	34	ambulance drivers	5
8	War posters	17	35	Food Administration	5
9	Women's Army Corps	16	36	hats	5
10	American red cross	14	37	Patriotism	5
11	France	14	38	Women in World War II	5
12	New York	14	39	women's history	5
13	African Americans	13	40	american flag	4
14	women's army auxiliary corps	13	41	Bermondsey	4
15	women workers	13	42	California	4
16	farming	10	43	civil war	4
17	gas mask	10	44	coast guard	4
18	flag	9	45	food conservation	4
19	Munitions	9	46	homefront	4
20	British	9	47	Indiana	4
21	factory	8	48	machine guns	4
22	feminism	8	49	Marine Corps	4
23	recruiting	8	50	Massachusetts	4
24	red cross	8	51	national history day	4
25	uniforms	8	52	spars	4
26	suffragists	7	53	Washington D. C.	4
27	welding	7	合计	—	702

表 3 标签共现矩阵(部分)

	African Americans	ambulance drivers	american flag	American red cross	Bermondsey	British	California	civil war	coast guard	factory	farming	feminism	flag	food
African Americans	13	0	0	0	0	1	2	0	0	0	0	0	1	0
ambulance drivers	0	5	0	0	0	0	0	0	0	0	0	0	0	0
american flag	0	0	4	0	0	0	0	0	1	0	0	1	4	0
American red cross	0	0	0	14	0	0	0	1	0	0	0	0	0	1
Bermondsey	0	0	0	0	4	1	0	0	0	4	0	0	0	0
British	1	0	0	0	1	9	0	0	0	0	0	0	0	0
California	2	0	0	0	0	0	4	0	0	0	0	0	0	0
civil war	0	0	0	1	0	0	0	4	0	0	0	0	0	0
coast guard	0	0	1	0	0	0	0	0	4	0	0	0	0	0
factory	0	0	0	0	4	0	0	0	0	8	0	0	0	0
farming	0	0	0	0	0	0	0	0	0	0	10	1	0	0
feminism	0	0	1	0	0	0	0	0	0	0	1	8	1	0
flag	1	0	4	0	0	0	0	0	0	0	0	1	9	0
food	0	0	0	1	0	0	0	0	0	0	0	0	0	6

聚类分析，可直观地看到 women、World War Ⅰ、World War Ⅱ、posters4 个大节点为本例的 4 个主题类型，与各主题类型有关联的高频标签为其所含主题，见图 4 左侧。

（2）同理，将标签共现矩阵导入到 NodeXL 中，运用 NodeXL 聚类功能并选择相应聚类算法进行聚类分析，将标签集聚类为标签群，得到图 4 右侧所示的 4 个标签共现关系类团，可以看出，其主题类型亦为 women、World War Ⅰ、World War Ⅱ、posters。综上，将 women、World War Ⅰ、World War Ⅱ、posters 这 4 个关键标签作为“Women at War”这个 tagging mission 里标签的主题类型。

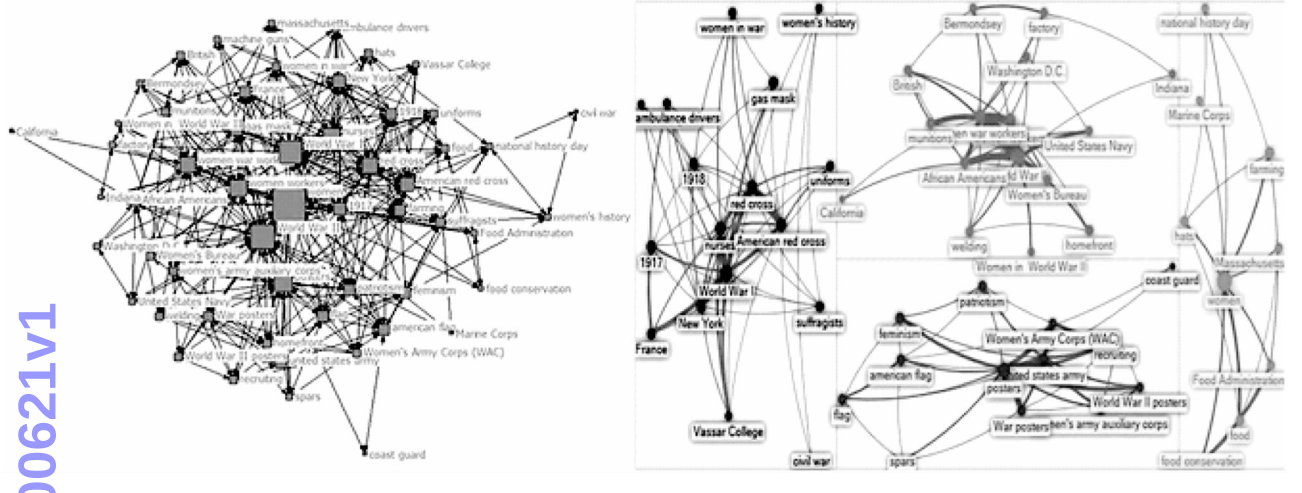


图 4　“Women at War”聚类结果对比分析

3.2.2　关联及关联类型分析　该阶段旨在分析所得的主题类型及主题间的关联关系与关联类型。以“posters”主题类型及其所涉及的主题为例，利用形式概念分析理论，将其所含主题集－资源集填充到二元表中，以主题为形式概念的内涵，以档案资源为概念的外延，以“x”代表主题－资源的二元关系，构建形式背景，然后利用概念格构造工具(conexp1.3)将上述形式背景转化为相应的概念格 Hasse 图，对主题实现层次化的聚类，见图 5。对该 Hasse 图中内涵和外延进行分析，总结归纳出 3 种关系：属分关系、包含关系、相关关系。举例来说，顶层的 posters 作为主题类型包含了所有主题，这就是包含关系；主题 flag 和主题 American flag 体现了形式概念的上下位继承关系，则主题 American flag 是主题 flag 的子主题，体现主题间的属分关系；主题 united states army 和主题 women's army corps 则是相关关系。其他关联关系就不一一赘述了。

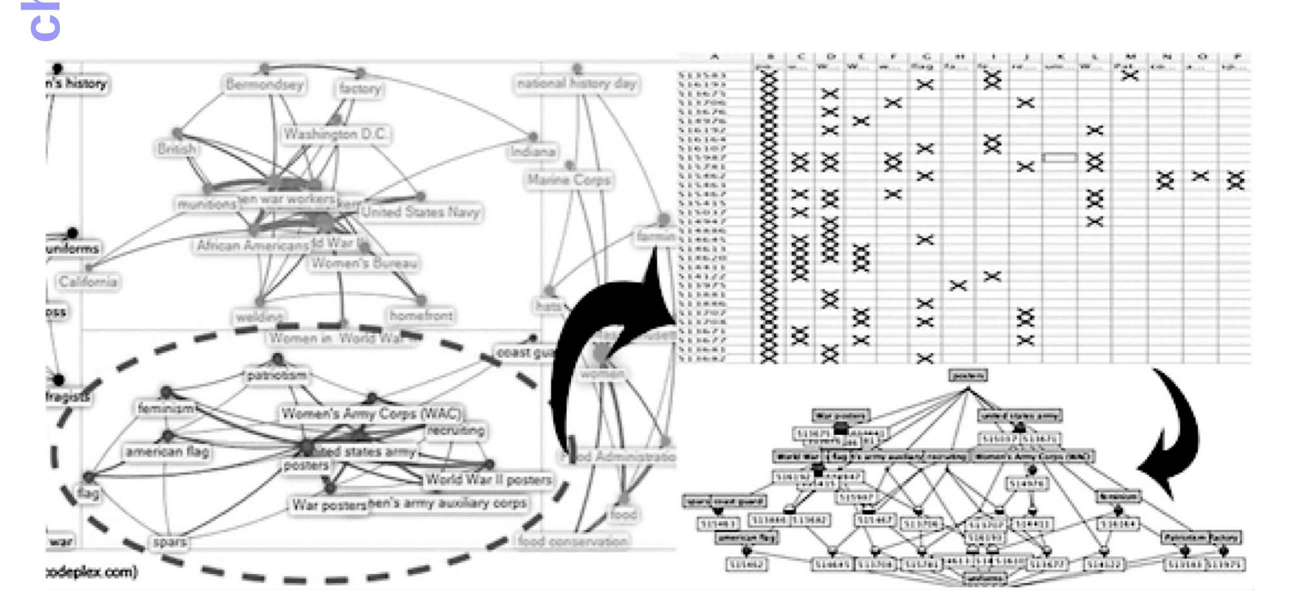


图 5　主题类型 posters 及其主题的 Hasse 图

3.2.3 资源指引分析 本阶段将在上一步的基础上进行分析，在相应的主题下链接资源实体。仍以主题类型“posters”中的主题“flag”和主题“american flag”为例，根据前文分析结果，主题“american flag”的链接资源应为 515462、514947、513673、533765 共 4 件案卷，而主题“flag”的链接资源应为 31488352、26432783、6788430、533657、535600、515462、514947、513673、533765 共 9 件案卷，结合本文所用形式概念分析理论可知，其中后 4 件案卷可视为从主题“american flag”处逆向继承得来。采用这种聚合式资源指引方式后，若以“flag”为检索词，其返回 9 项结果中，案卷 31488352、26432783、6788430、533657、535600 的排序应优先于其他 4 件案卷。

3.3 基于 Ontopia 创建关于 NARA 数字档案标注系统的主题图

本阶段利用 OKS 中的主题图编辑器 ontopoly、浏览器 Omnigator、可视化(Ontopia Navigator)工具进行主题图的编辑、浏览与可视化，实现数字档案标注系统的资源聚合。

3.3.1 利用 Ontopoly 创建主题图 Ontopoly 分为本体(ontology)编辑器和实例(instances)编辑器两部分，本阶段先通过 Ontopoly 界面的类型索引页和类型配置页对“Women at War”的主题及主题类型、关联及关联关系和资源指引进行本体内容的编辑，然后用实例编辑器对各主题的实例进行编辑输入，从而实现主题图的创建，其结果如图 6 所示：

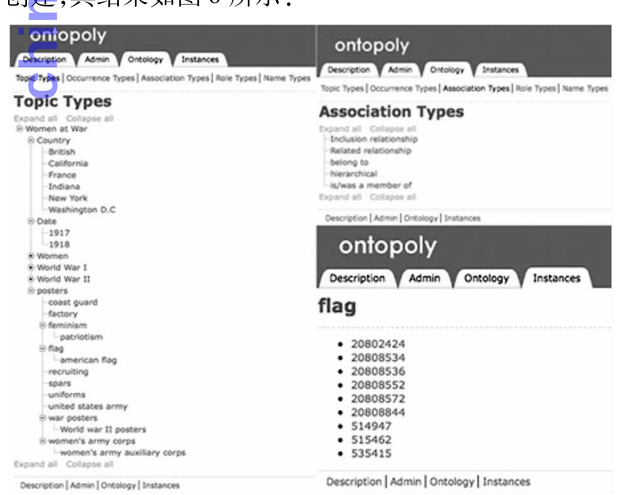


图 6 Ontolopy 创建的“Women at War”主题图

(1)使用 topic types 模块创建主题及主题类型，将 3.2 中分析出来的内容特征主题类型“women”“World War I”“World War II”“posters”和外部特征主题类型如年份、国家/地区等输入到该模块中并添加其相应的主题，在其主题配置和主题类型配置页面设置各自的属性；然后再使用实例(instances)编辑器对各主题的实例进行编辑输入，如为主题 flag 添加对应的实例有国家档案馆标识号为 20808534、20808572、20808844、20808536、20802424、20808552 等案卷。

(2)使用 Association Types 模块创建关联及关联关系。可描述的关联关系涵盖包含关系、属分关系、相关关系等。以 Posters 中的 flag 为例，这个概念的下位概念有 american flag，可在该模块中对其进行“属分关系”的编辑，其他的关系也可参照如此编辑。

(3)使用 Occurrence Types 模块创建与对应主题相关的资源属性和资源类型，见表 4。例如，可在主题 flag 的类型配置页面添加资源属性：简介、资源来源、类型名称、代表及含义，还可为其添加资源类型如 HTML、Image 及相关的资源链接。

表 4 资源指引属性及其类型

主题类型	资源属性	资源类型
Country	简介、资源来源	HTML、URI、String、Image
Date	简介、资源来源、代表、	Number、Datetime
posters	简介、资源来源、类型名称、代表、含义	HTML、URI、Number、Datetime、String、Image
women	简介、资源来源、类型名称、代表、含义	HTML、URI、Number、Datetime、String、Image
World War I	简介、资源来源、类型名称、代表、含义	HTML、URI、Number、Datetime、String、Image
World War II	简介、资源来源、类型名称、代表、含义	HTML、URI、Number、Datetime、String、Image

3.3.2 利用 Omnigator 浏览主题图 基于主题图的资源聚合与导航结果可通过 Omnigator 的主页面来展示，Omnigator[11] 浏览器是一个标准的 Web 界面，用户可直接浏览主题、关联关系、资源指引及其指引所给出的链接，并通过点击链接，到达相应的信息资源，从而将内部的主题、关联等和信息资源联系起来，见图 7。该浏览界面以文本的方式显示了“Women at War”中“posters”关联类型和主题实例等。点击图中的 Subject Identifiers，可以链接到该“posters”标签所对应的网页。

3.3.3 利用 Navigator 实现主题图的可视化 主题图可视化是指用一个表达语义的网状结构来描述主题与主题间关联关系。图 8 是由 Ontopia Visual Navigator 可视化组件生成的，以网状图的结构展示 NARA 数字档案标注资源间固有的和潜在的知识结构。图中每个主题上都有相关的数字，反映的是与该主题所关联的主题，例如主题 flag 右上方的 2 表示主题 flag 有两个相关联的主题，即主题类型 posters 和主题 american flag，

posters Type(s): Topic type

Name (1)
- posters

Associations (17)
- Has cardinality
  - Cardinality: Exactly one
  - Field definition: [No name]
- Has field
  - [No name]
  - [No name]
  - contain
  - HTML
  - The Battle for Fuel
- Subclass of
  - Women at War
- Superclass of
  - coast guard
  - factory
  - feminism
  - flag
  - recruiting
  - spars
  - uniforms
  - united states army
  - war posters
  - women's army corps

Subject Identifiers (1)
- https://catalog.archives.gov/search?q=*:*&f.oldScope=(descriptions%20or%20online)&f.tagsKeywordsAdv=posters&SearchType=advanced

Hierarchy (Superclass/subclass)
- Women at War
  - posters

Internal Occurrences (1)
- Field order
  - 000000001 - Scope: [No name]

Topics of this Type (13)
- Telephone Line Time is Precious - Keep Your Talks Short
- Telephone Line Time is Precious - Keep Your Talks Short
- Telephone Line Time is Precious - Keep Your Talks Short
- Ten Million New Members by Christmas - Join the Red Cross
- Ten Million New Members by Christmas - Join the Red Cross
- Ten Million New Members by Christmas - Join the Red Cross
- The ATS Wants Cooks
- The Attack Begins in the Factory [1]
- The Attack Begins in the Factory [2]
- The Battle for Fuel
- The Battle for Fuel - Turn the Gas Down
- The Battle for Fuel - Use Less Water
- The Battle for Fuel - Use Less Water

NATIONAL ARCHIVES CATALOG

Advanced Search Login Register

Register of Patients at St. Elizabeths Hospital

Cabell, General Charles P.

Put Your Trust in the Navy and Your Money in Savings Certificates

"It's smart to salt away some of your pay"

图7　“World War I”主题类型浏览界面

用户可以根据需要选择主题进一步追踪查询，能提高检索过程中的查准率和查全率。主题图不仅能可视化地展示整个主题及主题类型，还能具体到关联及关联关系，甚至是每个关联关系所链接的主题。

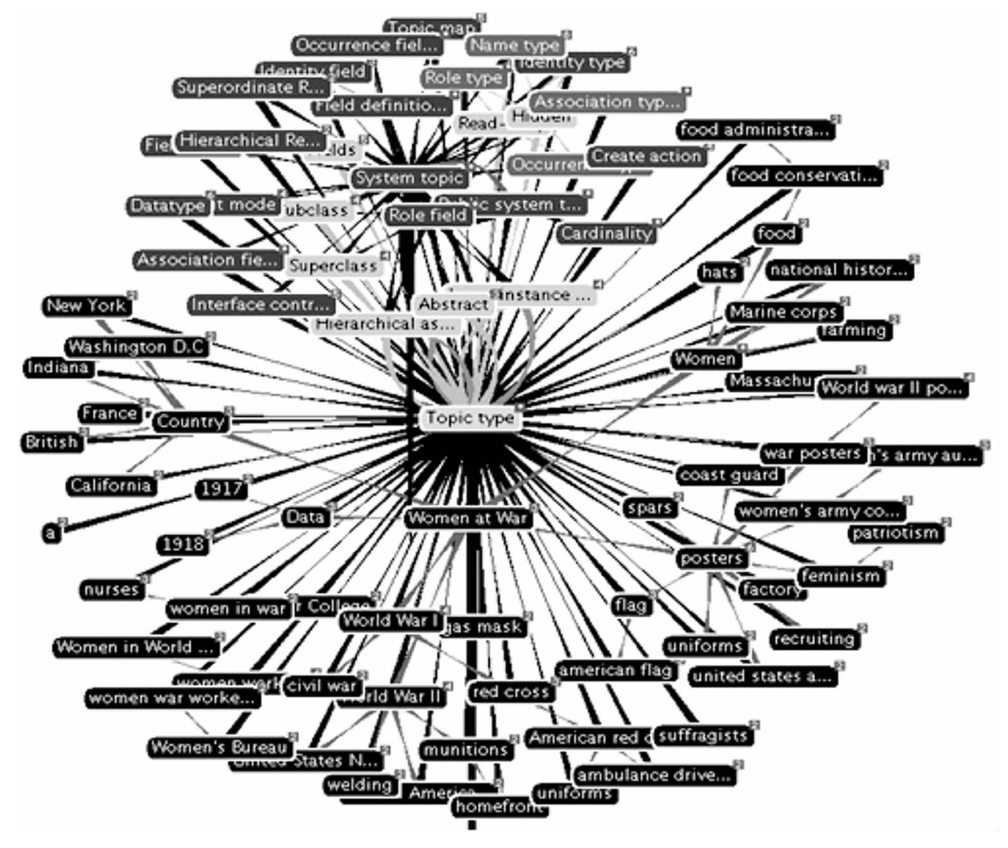


图8　“Women at War”主题类型可视化

4 结语

基于主题图实现资源聚合已成为学界诊治社会化标注系统中基于标签匹配实现资源检索与导航系列痼疾的一剂良药,本研究以 NARA 数字档案标注系统的资源聚合为例,重构了社会化标注系统三元组向主题图三要素之间的映射方案,给出了基于主题图的社会化标注系统资源聚合模型,采用社会网络分析、形式概念分析等量化工具更加科学严谨地建立了从社会化标注系统三元组到主题图三要素的映射,从而确保基于主题图的数字档案资源聚合结果能描述更精准的语义和展示更精确的导航。

本研究也存在诸多不足,例如因无法自动化地批量获取 NARA 数据集造成样本容量较小;侧重从数字档案资源内容特征的角度建立主题图而忽略了档案资源外部特征相关主题的刻画;再如缺少对基于主题图的数字档案资源聚合效果的定量评价等,这些将都是本研究后续要解决的问题。

参考文献:

[1] RATH H. Topic maps: templates, topology, and type hierarchies[J]. Acoustics speech & signal processing newsletter IEEE, 2000(2):45-64.

[2] 陈婷,胡改丽,陈福集,等. 社会化标注环境下的数字图书馆知识组织模型研究——基于标签主题图视角[J]. 情报理论与实践,2015,38(3):63-70.

[3] 胡娟,程秀峰,叶光辉. 基于主题图的学术博客知识组织模型研究[J]. 图书情报工作,2012,56(24):127-132.

[4] 项兴彬. 建筑企业知识标签主题图构建研究[J]. 信息系统工程,2016(4):96-97.

[5] Cleaning the skies: from tag clouds to topicmaps[EB/OL]. [2017-04-29]. http://www.topicmaps.com/tm2007/lavik.pdf.

[6] HENDEL D, KUZHABEKOVA A, CHAPMAN W. Mapping global research on international higher education[J]. Research in higher education, 2015,56(8):861-882.

[7] FUJIMURA K, IWATA T, HOSHIDE T, et al. Geo topic model: joint modeling of user's activity area and interests for location recommendation[C]// ACM international conference on web search & data mining. New York:ACM,2013:375-384.

[8] 熊回香,邓敏,郭思源. 标签主题图的构建与实现研究[J]. 图书情报工作,2014,58(7):107-112.

[9] 邓敏. 基于主题图的标签语义挖掘研究[D]. 武汉:华中师范大学,2014.

[10] 夏立新,张玉涛. 基于主题图构建知识专家学术社区研究[J]. 图书情报工作,2009,53(22):103-107.

[11] Omnigator:the topic map browser[EB/OL]. [2017-05-05]. http://www.ontopia.net.

作者贡献说明:

张云中:论文选题、策划框架和设计方案,论文主体部分撰写并指导修改;

冯双双:收集与整理文献,收集数据,开展实验,撰写论文初稿并修改论文。

Study on Resource Aggregation of Digital Archives Tagging System Based on Topic Maps

Zhang Yunzhong Feng Shuangshuang

School of Library, Information Science and Archive, Shanghai University, Shanghai 200444

Abstract: [**Purpose/significance**] Aiming at the problems of resource retrieval and navigation which are caused by social tagging system used for digital archive resource organization, a digital archive resource aggregation model based on topic maps is presented in order to improve the efficiency of digital archive resource retrieval and establish an orderly visual navigation. [**Method/process**] Based on the analysis of using topic maps to realize the research status of social tagging system resource aggregation, a resource aggregation model based on topic maps in the field of digital archives is constructed, and a systematic solution to the three key elements of digital archives resource: topic types, association types and occurrence types is given, which uses social network analysis and formal concept analysis, so as to realize the resource aggregation of digital archives tagging system. [**Result/conclusion**] Taking the topic of "Women at War" in NARA digital archives tagging system as an example, we use the method proposed in this paper and combine Ontopia tools to achieve the aggregation of digital archives resources of target topic, which effectively improves the retrieval efficiency and navigation effect of digital archives resources.

Keywords: social tagging systems topic maps resource aggregation digital archives